DOCUMENT RESUME

ED 471 226 TM 034 659

AUTHOR Kester, Donald L.; Linton, Thomas H.; Sullivan, Lynn R.

TITLE A Comparison of the Relative Practical Value of a Predictive Discriminant Function Analysis and a Binary Logistic Regression Analysis of Student Success in an Innovative Alternative High School Program in South Texas: Data Set #1. Working Paper.

PUB DATE 2002-04-00

NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC01 Plus Postage.

DESCRIPTORS High Risk Students; *High School Students; High Schools; *Prediction; Regression (Statistics); *Student Characteristics

IDENTIFIERS *Logistic Regression; *Predictive Discriminant Analysis; Wisconsin

ABSTRACT

This study compared two statistical approaches, predictive discriminant function analysis and binary logistic regression analysis, to determine the characteristics of at-risk students who succeeded in comparison with those who did not succeed in a project-based alternative high school program. The study also investigated true positive hit rates for predicting correctly the students that would be successful using each statistical approach to provide information for decision making in admissions decisions for this program. Data came from the Wisconsin Youth Survey from measures of bonding with teachers, school, and peers and from a measure of educational engagement completed by 70 students. Other variables measured were family composition, student employment, grade, gender, and age. The predictive discriminant analysis approach produced a true positive hit rate of 79% compared to the remarkable 94% of the binary logistic approach. Advantages of the binary logistic approach are discussed. (SLD)

A Comparison of the Relative Practical Value of a Predictive Discriminant Function Analysis and a Binary Logistic Regression Analysis of Student Success in an Innovative Alternative High School Program in South Texas

Data Set # 1

A Working Paper

by

Donald L. Kester, Ph.D.

and

Thomas H. Linton, Ph.D.

Texas A&M University – Corpus Christi

and

Lynn R. Sullivan, Ed.D.

Corpus Christi Independent School District

Corpus Christi, TX

This Paper was presented at the American Educational Research Association's Annual Conference held in New Orleans, LA during April 1-6, 2002.

ABSTRACT

A. Purposes: 1) to compare two statistical approaches; i.e., predictive descriminant function analysis and binary logistic regression analysis to determine the characteristics of at-risk students who succeeded with those who did not succeed in a project based alternative high school program, 2) to compare true positive hit rates for correctly predicting which students would be successful using each statistical approach (SPSS, 1999a, 1999b), and 3) to develop one or more decision making strategies for use in admissions decisions for this volunteer program.

B. Theoretical Framework: The primary conceptual base for this action research/program evaluation study was the dropout prevention theory developed by Wehlage, Rutter, Smith, Lesko, & Fernandez (1989). This Theory includes an emphasis on the importance of a student's social bonding, especially bonding to teachers, to school, and to peers.

C. Methods: Both statistical approaches were used to predict membership in the successful student category of the dichotomous ("successful" or "not successful') criterion variable. The same data set and nine predictor variables were used in both analyses.

D. Data Sources: A primary data source was the Wisconsin Youth Survey's scales of Social Bonding to Teachers, Social Bonding to School, and Social Bonding to Peers (Wehlage, Rutter, & Stone, 1986). In addition, Sullivan (2000) created an Educational Engagement scale to measure each student's perception of the alternative high school curriculum. Other variables measured were family composition, student employment, grade, gender, and age, each of which had been mentioned as relevant in earlier research on dropout prevention. The operational definition of the criterion variable "student success" was based on attendance, course completion, and disciplinary referrals.

E. Results: The predictive discriminant analysis approach produced a true positive hit rate of 79% (79.2%) compared to the binary logistic approach's remarkable 94% (93.8%). By definition, the term "true positive hit rate" refers to the accuracy of predicting membership in the successful student category for the sample of 70 students.

F. Educational Importance: Using the binary logistic approach is seen as advantageous in several ways: 1) the true positive hit rate was appreciably higher for the sample, 2) it does not require the

distributional assumptions that predictive discriminant analysis does (Tabachnick & Fidell, 2000), and 3) it can lead to an easily understood probability statement for an individual student (SPSS, 1999b). Binary logistic is therefore recommended for use either by itself or in combination with PDA for this particular program. Further research, including replication and detailed examination of cases which fell into the "false positive" and "false negative" classifications is recommended.

A Comparison of the Relative Practical Value of a Predictive Discriminant Function Analysis and a Binary Logistic Regression Analysis of Student Success in an Innovative Alternative High School Program in South Texas

INTRODUCTION

In her dissertation, Sullivan (2001) used nine predictor variables and predictive discriminant function analysis (PDA) to categorize at-risk students (N=70) attending an alternative high school program into either the "successful" student group or the "unsuccessful" student group. The predictor variables she used were: "age, gender, student employment, family composition, grade level, social bonding to teachers, social bonding to school, social bonding to peers, and students' perception of project based learning (educational engagement) (p.84)."

All predictors were either directly drawn from previous research, or were derived from the bonding-oriented, dropout prevention theory of Wehlage, Rutter, Smith, Lesko, and Fernandez (1989). The educational engagement scale was created by Sullivan (2001) to fit both the dropout prevention theory of Wehlage et al. (1989) and the unique setting of the particular south Texas alternative high school program involved in the study. The Wisconsin Youth Survey was used to obtain data on the students: Social bonding to

school, Social bonding to Teachers, and Social bonding to Peers (Wehlage, Stone, and Rutter, 1986; Wehlage, Rutter, and Turnbaugh, 1987). Data on remaining predictors were obtained via student survey.

As mentioned, the criterion variable of "School Success" was defined in terms of a student's attendance, number of credit completions, and number discipline referrals. More precisely, "Students who had less than five unexcused absences, two or more course completions, and less than four disciplinary referrals were placed in the success group (Sullivan, 2001, p. 22)."

Sullivan (2001) was especially interested in moving away from placing the blame for the dropout problem every where except with educational institutions. In her review of the literature she found "high divorce rates, broken homes, poverty, heredity, drugs, and other facts of society absorbed the blame of the problem. Blame was also placed on the dropouts themselves for their high absentee rates, lack of motivation, and wanton behavior (U.S. Department of Education, 1994a) (p. 1)."

She was interested in discovering those things educators could change within educational institutions to lower the dropout rate. She chose the predictive discriminant analysis (PDA) approach to help solve the problem.

But PDA is only one of two statistical approaches that could be used. Binary Logistic Regression is the other.

Competing Statistical Approaches

Binary Logistic regression might be preferable. Let's examine both approaches, beginning with PDA.

Predictive Discriminant Analysis:

Stevens (1996) states, "Recall that in multiple regression we found the linear combination of the predictors that was maximally correlated with the dependent variable. Here in descriminant analysis linear combinations are again used to distinguish the groups. (p. 262)."

But Field (2000) observes, "...there is a good reason why we cannot apply linear regression directly to a situation in which the outcome variable is dichotomous... for linear regression to be a valid model, the observed data should contain a linear relationship. When the outcome variable is dichotomous, this assumption is usually violated (see Barry, 1993) (p. 165)."

One reason why logistic regression may be better, is that one or more of the assumptions on which predictive discriminant analysis is based, may be violated.

To review, the following three assumptions underlie significance testing in predictive discriminant analysis:

> Assumption 1: The Quantitative Variables Are Multivariately Normally Distributed for Each of the Populations, with the Different Populations Being Defined by the Levels of the Grouping Variable.
>
> Assumption 2: The Populations Variances and Covariances among the Dependent Variables Are the Same across All Levels of the Factors.
>
> Assumption 3: The Participants Are Randomly Sampled, and the Score on a Variable for Any One Participant Is Independent from the Scores on This Variable for All Other Participants (Green, Salkind and Akey, 2001, pp. 279, 280).

Green et al. (2001) further noted, "If the dependent variables are multivariately normally distributed, each variable is normally distributed ignoring the other variables, and each variable is normally distributed at every combination of values of the other variables…it is hard to imagine that we would ever meet this assumption…(and) to the extend that the sample sizes are disparate and the variance and covariance's are unequal, the p-values yield invalid results (p.280)." Another problem with predictive discriminant analysis is that it "…is highly sensitive to the inclusion of outliners (Tabachnick & Fidell, 2001, p. 462)."

Although, as one set of authors write, "… it is reassuring that discriminant analysis yields relatively valid results in terms of Type I errors with moderate to large sample sizes (Green, Salkind, & Akey, 2001, p. 280)," the sample sizes in the current study may not be big enough. Out of 70 at-risk students, 48 were "successful" and 22 were "unsuccessful." Tabachnick &

Fidell (2001) write, "As a conservative recommendation, robustness is expected with 20 cases in the smallest group if there are only a few predictors (say, five or fewer) (p. 462)." But the current study does not have "five or fewer" predictors, it has nine.

Logistic Regression:

Tabachinick & Fidell (2001) describe this approach as follows:

> Logistic regression allows one to predict a discrete outcome such as group membership from a set of variables that may be continuos, discrete, dichotomous, or a mix (Emphasis added). Because of its popularity in the health sciences, the discrete outcome in logistic regression is often disease/no disease. For example, can presence or absence of hay fever be diagnosed from geographic area, season, degree of nasal stiffness, and body temperature?
>
> Logistic regression is related to, and answers the same questions as, discriminant function analysis (Emphasis added). the logit from of multiway frequency analysis with a discrete DV (Dependent Variable), and multiple regression analysis with a dichotomous DV. However, logistic regression is more flexible than the other techniques. Unlike discriminant function analysis, logistic regression has no assumptions about the distributions of the predictor variables; in logistic regression, the predictors do not have to be normally distributed, linearly related or of equal variance within each group (Emphasis added).
>
> There may be two or more outcomes in logistic regression. If there are more than two outcomes, they may or may not have order (e.g., no hay fever, moderate hay fever, severe hay fever). Logistic regression emphasizes the probability of a particular outcome for each case (Emphasis added). For example, it evaluates the probability that given person has hay fever, given that person's pattern of responses to questions about geographic area, season, nasal stuffiness and temperature.
>
> Logistic regression analysis is especially useful when the distribution of responses of the DV is expected to be nonlinear with one or more of the IV's (Independent Variables) (Emphasis added). For example, the probability of heart disease may be little affected (say 1%) by a 10-point difference among people with low blood pressure (e.g., 110 vs. 120) but may change quite a bit (say 5%) with an equivalent difference among

> people with high blood pressure (e.g., 180 vs. 190). Thus, the relationship between heart disease and blood pressure is not linear (p. 517).

Results

Which of two statistical approaches –predictive discriminant analysis and binary logistic regression – was better at identifying those at risk students who would be successful in the alternative high school program?

Using an available data set, the binary logistic analysis was better.

The true positive hit rate for the predictive discriminant analysis approach was only 79 %, as shown in Table 1 below.

Table 1

Predictive Discriminant Analysis Classification Results

Success (S)		Predicted Group					
Non-success (NS)		Success		Non-Success		Total	
Original	(S)	38	(79.2%)	10	(20.8%)	48	(100%)
	(NS)	4	(18.2%)	18	(81.8%)	22	(100 %)
	Totals	42		28		70	
Cross-validated (S)		33	(68.8%)	15	(31.3%)	48	(100%)
	(NS)	8	(36.4%)	14	(63.6%)	22	(100 %)
	Totals	41		29		70	

The comparable hit rate for the binary logistic analysis—using the same data set—was much higher, 94 %, as shown in Table 2.

Table 2

Binary Logistic Analysis Classification Results

	Predicted Group					
	Success		Non-Success		Total	
Success (S)	45	(93.8%)	3	(6.2%)	48	(100%)
Non-success (NS)	8	(63.6%)	14	(36.4%)	22	(100 %)

Conclusion:

Using the binary logistic approach is seen as advantageous in several ways: 1) the predictive true positive hit rate was appreciably higher for the sample, 2) it does not require the distributional assumptions that discriminant analysis does (Tabachnich & Fidell, 2000), and 3) it can lead to an easily understood probability statement for an individual student (SPSS, 1996b). Binary logistic is therefore recommended for use either by itself or in combination with predictive discriminant analysis for this particular program.

Discussion:

Ideally, the comparison of the accuracy of prediction by the two statistical techniques could be replicated on several new populations of students. Even if this took place only in this particular alternative high school program in south Texas, these replications would help answer the question of how typical are the results found here. Would the true positive hit rate for logistic analysis continue to be higher by almost 15 %, and would it remain in the mid – 90% range?

Another topic for future investigation could be a closer look at those students predicated to be successful, but who were, in fact, unsuccessful; i.e., the "false positives." The number appears to be small enough to allow for a detailed examination of only a few student records. For example, did most of these incorrectly categorized students barely miss qualifying as successful? By definition, to be "successful" a student must have had: (1) less than five unexcused absences, (2) two or more course completions, and (3) less than four disciplinary referrals. Of the false positives for example, how may missed being "successful' because they had five or six unexcused absences but otherwise met the definition of "successful?" In other words, to what degree and in what ways were the predictions inaccurate? The same examination could take place for the "false negatives;" that is, those students

predicted to be "unsuccessful" who turned out to be instead "successful." Such detailed case study information might help in making decisions as to which students should be admitted to this program in the future.

References

Field A. (2000). Discovering statistics using SPSS for Windows: Advanced techniques for the beginner. Thousand Oaks, CA: Sage.

Green, S.B., Salkind, N.J. , and Akey, T.M. (2000). Using SPSS for Windows: Analyzing and understanding data. (2nd Ed.) Upper Saddle River, NJ: Prentice Hall.

Stevens, J. (1996). Applied multivariate statistics for the social sciences. (3rd Ed.) Mahwah, NJ: Lawrence's Erlbaum Associates.

Sullivan, L.R. (2000). Predictors of success and non-success of at-risk secondary students at a project-based alternative high school (doctoral dissertation, Texas A&M University at Corpus Christi and Texas A&M University at Kingsville, 2000).

SPSS, Inc. (1999a). SPSS Base 9.0 application guide. Chicago: SPSS Inc.

SPSS, Inc (1999b). SPSS Regression Models 10.0. Chicago: SPSS, Inc.

Tabachnick, B.G., & Fidell, L.S. (2000). Using multivariate statistics, (4th Ed.) Boston: Allyn and Bacon.

Wehlage, G.G., Rutter R.A., Smith, G.A., Lesko, N., N., & Fernandez, R.R. (1989). Reducing the risk: Schools as communities of support. New York: Falmer.

Wehlage, G.G., Rutter, R.A., & Stone, C. (1986). Wisconsin Youth Survey, Madison, WI: National Center on Effective Secondary Schools.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM034659

Reproduction Release
(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title:
A Comparison of the Relative Practical Value of a Predictive Discriminate Function Analysis and a Binary Logistic Regression Analysis of Student Success in an Innovative Alternative High School Program in South Texas Data Set #1

Author(s): Donald L Kester, Ph.D., Thomas H Linton, Ph.D. and Lynn Sullivan, Ed. D.

Corporate Source: Texas A&M University – Corpus Christi
6300 Ocean Drive
Corpus Christi, TX 78412

Publication Date:
April 6, 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	**Level 2A**	**Level 2B**
↑ X	↑ ☐	↑ ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature:	Printed Name/Position/Title: Dr. Don Kester	
Organization/Address: Texas A&M University – Corpus Christi ean Drive hristi, TX 78412	Telephone: (361) 825-2175	Fax: (361) 825-2732
	E-mail Address: Don.Kester@mail.tamucc.edu	Date: November 11, 2002

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706
Telephone: 301-552-4200
Toll Free: 800-799-3742
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org

EFF-088 (Rev. 2/2001)